Hand Tracking Using Optical-Flow Embedded Particle Filter in Sign Language Scenes

Selma Belgacem[1], Clément Chatelain[1], Achraf Ben-Hamadou[2], and Thierry Paquet[1]

[1] LITIS EA 4108, University of Rouen
Saint-Etienne du Rouvray, France
selma.belgacem@etu.univ-rouen.fr
[2] University of Paris-Est, LIGM (UMR CNRS), Center for Visual Computing, ENPC
Marne-la-Vallée, France
achraf.ben-hamadou@ensem.inpl-nancy.fr

Abstract. In this paper we present a method dedicated to hand tracking in sign language scenes using particle filtering. A a new penalisation method based on the optical flow mechanism is introduced. Generally, particle filters require the use of a reference model. In this paper we have introduced a new method based on a dictionary of visual references of hand to constitute the reference model. The evaluation of our method is performed on the SignStream-ASLLRP database on which we have provided ground truth annotations for this purpose. The obtained results show the accuracy of our method.
Keywords: hand tracking, particle filtering, optical flow, hand vocabulary, sign language scene.

1 Introduction

In this article we propose a method for hand tracking in sign language scenes. Hand gestures are characterised by frequently changing hand configuration (fingers and palm pose) and random motion [13], thus requiring a robust and accurate tracking method.

Particle filter [3, 10] is a state-of-the-art framework based on a probabilistic *predictive* tracking formalism that has shown to be efficient in various applications as sports tracking [6, 16], face and hand tracking [1, 4, 11] and vehicle tracking [5]. Prediction is based on a Markovian motion model, and an iterative Monte Carlo *weighted Sampling* applied on a set of particles. *Particles* are the target region hypotheses, typically points, bounding boxes or more complex geometrical models. Particle filters are based on three essential models: *observation model* which weights particles according to the associated extracted measurements, *reference model* which is a reference representation of the tracked object, and *motion model* according to which particles are propagated. In this paper, we present a contribution to each of these stages which we introduce in the *condensation* implementation [4] of a particle filter.

Our main contribution is the integration of estimated and observed motion information in the motion model and the observation model. Indeed, the random aspect of hand gestures in sign language scenes makes it difficult to use predefine motion models. In this respect, Bhandarkar et al. and Yao et al. [1, 16] have introduced an optical-flow-based velocity term to the classic equation of the particle filter motion model. Optical-flow technique is known for its robustness against luminosity variations and deformations of the tracked object shape [1]. In the case of multiple objects moving in the same sequence, this observed velocity term becomes ambiguous. We propose to integrate similar information in the motion model based on the estimated position provided by the filter and weighted by a global observation deduced from optical-flow. The dominant hand in sign language has mostly the dominant motion in the scene. Then, a global velocity observation is highly influenced by dominant hand motion. Optical-flow observation can also be exploited locally to enhance the observation model. In fact, particles which move against the observed flow should be penalised. We propose a new method to apply this optical-flow local penalization by *re-weighting* particles.

In the framework of particle filtering, particles *weights* are iteratively computed using the *observation likelihood*. This observation likelihood is generally estimated thanks to a distance between a particle associated observation and the reference model. The reference model can be determined by either an initial detection [15] or an off-line learning process [9]. The first strategy is highly sensitive to deformations of patterns while the second requires an annotated data. We design a new method to automatically build a reference model. It is based on the construction of a vocabulary of the tracked object images thumbnails (figure 1) with different configurations, collected from the sequence in which the object will be tracked afterwards. In addition, our observation model is based on features invariant to deformation.

The outline of our paper consists of three sections. Section 2 introduces our observation and reference models. Section 3 explains our motion model and optical-flow penalisation at the global and local levels. Section 4 presents the experiments conducted and the evaluation results.

2 Particle filter

In this study, a particle X^i ($i \in \{1, \ldots, \mathcal{N}\}$) is associated to a bounding box defined by $p^i = (x, y)$ the particle position in the image and $s^i = (w, h)$ its width and height. $\mathcal{N}$ is the number of particles. A weight π^i is associated to each particle and is proportional to the observation likelihood $P(Y^i|X^i)$ where Y^i is a *features vector* associated to a particle X^i. Finally, for a given frame t, the *estimated position* $\hat{X}_t$ of the target $\mathtt{T}$ is the barycentre of the set of particles. In our case, the target $\mathtt{T}$ is the right hand.

Next in this section, we detail our reference and observation models involved in the particle filter.

2.1 Reference model : hand vocabulary

Since the hand is a deformable object, its appearance changes very often in the images. We, therefore, chose to use a *vocabulary* of hand appearances as a reference model (see Figure 1).

Fig. 1. Sample from hand vocabulary automatically extracted from a sequence

This vocabulary is built automatically off-line from the video sequence `S` as follows. We first use the well-known and robust face detection method of Viola and Jones [14] to localise the face in the first image of `S`. This allows for extracting a prior information about the colour range (*i.e.,* histogram) of the skin. Then, we extract skin blobs from the whole images using histogram back-projection and CamShift [2] algorithms. Afterwards, using some geometric assumptions, we select the most likely blobs standing for the right hand which is our target object `T`. It is worth noticing that we do not retain ambiguous configurations such as hand intersections or when the right hand is very close to the face. Finally, we end up with a set of cropped images of the hand to be tracked over the sequence `S`. Note that this method of skin blob detection is also used to localise the hand in the first image of `S` and initialize the filter.

The set of cropped images represent our hand vocabulary. We associate a reference feature vector Y^R to this vocabulary. Y^R is the average of the feature vectors of all the cropped images.

2.2 Observation model

The observation model allows the filter to compare a given particle X^i with the reference model so that a weight π^i is computed according to its similarity to the model.

Selected features. The most important features for hand tracking are colour and shape. The colour is a classic feature used in the observation model of particle filters for tracking. The hand is characterised by a skin colour range. In our case, the *skin colour histogram* is represented in the HSV colour space as is very often used [12].

In sign language scenes, colour features are not sufficient to discriminate between the hand and the face. Therefore, we additionally consider two complementary shape descriptors namely, Hu and Zernike moments. Hu moments are invariant with respect to translation, scaling and symmetry of shapes while Zernike moments are invariant with respect to rotation.

Observation likelihood. Following the Condensation algorithm, $\pi^i = P(Y^i|X^i)$ $\forall\ i \in \{1,\ldots,\mathcal{N}\}$. We compute these weights using equation (1) which has a simple form that we define.

$$P(Y^i|X^i) = \prod_{l=1}^{m} \left(\frac{1}{1 + D_l(Y^i, Y^R)} \right)^{c_l} \tag{1}$$

In equation (1), m is the number of features, $c_l \in \mathbb{R}^+$ is used to give importance to some features, D_l measures a distance for the feature l between the feature vector Y^i of a particle i and the feature vector Y^R of our reference model. There is a specific D_l for each feature l.

We present in the next section our motion model and optical-flow penalisation.

3 Optical flow penalisation

Our goal is to integrate in the particle filter information about motion. First, we compute an optical-flow map Ψ_t for each frame t of `S` using Lucas-Kanade method [7]. Then, the integration is done at two levels: particles weights and particles motion model.

3.1 Velocity and particles re-weighting

The idea here is to penalise particles which are moving against the observed flow. To do so, we characterise each particle X_t^i with $\boldsymbol{\nu}_t^i$ which is the median velocity computed from the corresponding X_t^i window in the Ψ_t map. The optical-flow penalisation of particles with $\boldsymbol{\nu}_t^i$ is done via a kind of weighting term ξ_t^i which we define as follows:

$$\xi_t^i = \frac{1}{1 + \lambda_t^i} [cos(\widehat{\rho\boldsymbol{\nu}_t^i, \dot{p}_t^i})]^{\tau_t^i}. \tag{2}$$

In the equation (2), $\dot{p}_t^i$ is the particle displacement vector, $\rho = \delta t$, and λ_t^i and τ_t^i values are defined in the table 1 according to conditions on optical-flow observation $\boldsymbol{\nu}_t^i$ and the associated particle displacement $\dot{p}_t^i$.

Table 1. (λ_t^i, τ_t^i) values according to conditions on optical-flow observation and the associated particle displacement

Conditions	condition 1	condition 2	condition 3	condition 4
	$\|\rho\boldsymbol{\nu}_t^i\| = 0$ AND $\|\dot{p}_t^i\| = 0$	$\|\rho\boldsymbol{\nu}_t^i\| = 0$ XOR $\|\dot{p}_t^i\| = 0$	$\|\rho\boldsymbol{\nu}_t^i\|\|\dot{p}_t^i\| \neq 0$	
			$cos(\widehat{\rho\boldsymbol{\nu}_t^i, \dot{p}_t^i}) \leq 0$	$cos(\widehat{\rho\boldsymbol{\nu}_t^i, \dot{p}_t^i}) > 0$
(λ_t^i, τ_t^i) values	$(0, 0)$	$(\Lambda, 0)$	$(\Lambda, 0)$	$(\|(\|\rho\boldsymbol{\nu}_t^i\|, \|\dot{p}_t^i\|)\|_1, 1)$

In table 1, $|\ |_1$ stands for L_1-norm, $\Lambda \in \mathbb{R}^+$ is an empirical value which should be chosen big enough to make ξ_t^i tends to 0. In the case of condition 1, the particle X^i and the associated observed flow $\boldsymbol{\nu}_t^i$ are stationary, then X^i is not penalized. In the case of conditions 2 and 3, X^i and $\boldsymbol{\nu}_t^i$ have opposite states, then X^i is maximally penalized. In the case of condition 4, X^i and $\boldsymbol{\nu}_t^i$ have the same orientation, then X^i is only penalized by the velocity value difference $|(\|\rho\boldsymbol{\nu}_t^i\|, \|\dot{p}_t^i\|)|_1$ and the direction difference $cos(\widehat{\rho\boldsymbol{\nu}_t^i, \dot{p}_t^i})$.

ξ_t^i is then used to re-weight particles as follows: $\pi_t'^i = \pi_t^i \xi_t^i$, where $\pi_t'^i$ is the new weight of a particle X_t^i. Afterwards, particles are sampled according to $\pi_t'^i$. This first integration of optical-flow is qualified as local penalisation. Next, we present our motion model with an optical-flow global penalisation.

3.2 Velocity and particles motion model

The classic particle motion prediction equation according to the condensation algorithm is:

$$X_t^i = AX_{t-1}^i + BR_t^i. \tag{3}$$

In the equation (3), A is the transition matrix, R_t^i is a random vector and B is a random walk matrix. In our case, A and B are constant. As explained before, the signing hand motion model should be more elaborated to improve the tracking. Thus, we keep the classic prediction equation (3) and we introduce the velocity and acceleration of the filter estimation $\hat{X}_{t-1}$ as follows:

$$X_t^i = AX_{t-1}^i + BR_t^i + \alpha_t \begin{bmatrix} \dot{\hat{p}}_{t-1} \\ 0 \\ 0 \end{bmatrix} + \beta_t \begin{bmatrix} \ddot{\hat{p}}_{t-1} \\ 0 \\ 0 \end{bmatrix} \tag{4}$$

In equation (4), $\dot{\hat{p}}_{t-1}$ and $\ddot{\hat{p}}_{t-1}$ are respectively the displacement vector and the acceleration vector computed from the previous estimated positions, α_t and β_t are two 4×4 diagonal matrices gathering coefficients to weight the filter velocity and acceleration respectively. We define them as follows:

$$\alpha_t = \begin{pmatrix} \frac{\bar{\vartheta}(\Psi_t^x)}{\max_j \vartheta^j(\mathtt{S})} & 0 & 0\ 0 \\ 0 & \frac{\bar{\vartheta}(\Psi_t^y)}{\max_j \vartheta^j(\mathtt{S})} & 0\ 0 \\ 0 & 0 & 0\ 0 \\ 0 & 0 & 0\ 0 \end{pmatrix} \beta_t = \begin{pmatrix} \frac{\bar{\gamma}(\Psi_t^x)}{\max_j \gamma^j(\mathtt{S})} & 0 & 0\ 0 \\ 0 & \frac{\bar{\gamma}(\Psi_t^y)}{\max_j \gamma^j(\mathtt{S})} & 0\ 0 \\ 0 & 0 & 0\ 0 \\ 0 & 0 & 0\ 0 \end{pmatrix}$$

where $\vartheta^j(\Psi_t^x)$ is the absolute value of the x-axis velocity component for a pixel j (resp. y-axis), and $\gamma^j(\Psi_t^x)$ is the absolute value of the x-axis acceleration component for a pixel j (resp. y-axis). Taking into account both velocity and acceleration estimations in the motion model allows the generated particles to smoothly follow up T and to handle severe motion variations, respectively. By computing α_t and β_t from the whole velocity and acceleration maps, we handle the global motion in the scene. In fact, if the global motion is important in the scene, those coefficients will have important values, whereas, if the global motion is attenuated, those coefficients will have small values.

4 Experiments and results

4.1 Experiments

Evaluated systems. In order to assess the robustness of our method and to show the contribution of its components, we propose to compare four configurations of the particle filter, namely, `PF`, `VPF`, `2VPF`, and `3VPF`. `PF` is the classic particle filter using only the observation model presented in section 2.2. `VPF` is the `PF` with the use of a reference vocabulary. `2VPF` integrates the estimation velocity and acceleration to `VPF`. In that case, α_t and β_t have constant values determined experimentally. Finally, `3VPF` is the whole approach adding the optical-flow global and local penalisation to `2VPF`. The particle filter parameters are the same for the four systems, namely, $\mathcal{N}$ equals 100, A is the identity matrix, and B and c_l are experimentally determined.

Experimental data and evaluation criteria. We performed Hand tracking experiments on the American sign language database SignStream-ASLLRP [8]. It consists of four videos containing between 1310 and 5046 frames acquired in a recording studio. Their capturing rate is between 30 and 32 fps. Frames size is between 288×216 and 320×240. There is no constraints on signers clothes. We tuned our system parameters on $\mathtt{S}_1$, $\mathtt{S}_2$, and $\mathtt{S}_3$ video sequences, and we used $\mathtt{S}_4$ sequence for the evaluation. We build ground-truth data for these four videos by manually drawing a bounding box `G` on the dominant hand (the object to track) in all frames. From these drawings, we get for each frame, $\mathtt{G}_p$ the ground truth position of the hand and $\mathtt{G}_s$ its area. Our evaluation criteria are based on two measures: an error measure $\bar{\epsilon}$ (equation (5)) and the Jaccard index which is a ratio $\bar{\varrho}$ indicating the degree of overlap between the filter estimation $\hat{X}$ and the ground truth `G` (equation (6)). $\bar{\epsilon}$ measures `T` position tracking accuracy and $\bar{\varrho}$ measures `T` region tracking accuracy. $|S|$ stands for the number of frames in a sequence S.

$$\bar{\epsilon} = \frac{1}{|S|} \sum_{t=1}^{|S|} ||\hat{p}_t - \mathtt{G}_{p,t}|| \tag{5}$$

$$\bar{\varrho} = \frac{1}{|S|} \sum_{t=1}^{|S|} \frac{\hat{X}_t \cap \mathtt{G}_t}{\hat{X}_t \cup \mathtt{G}_t} \tag{6}$$

4.2 Results

Table 2 shows the results on the same $\mathtt{S}_4$ video according to $\bar{\epsilon}$ and $\bar{\varrho}$ measures. `PF` $\bar{\varrho}$ value is too small because the filter is totally distracted by the face and sometimes attracted by the hand when it passes closely. However, a clear progress is noticed between `PF` and `VPF`. This progress proves the contribution of our reference model. Table 2 shows also the contribution of our complete system `3VPF`, particle filter with optical flow penalisation. Clearly, it improves tracking performance. Moreover, figure 2 shows that the integration of velocity and acceleration

Table 2. Filter estimation position average error ($\bar{\epsilon}$) and matching average ratio ($\bar{\varrho}$) for $\mathtt{S}_4$:1310 frames

	PF	VPF	2VPF	3VPF
$\bar{\epsilon}$	54.28	31.82	29.23	21.22
$\bar{\varrho}$	0.004	0.279	0.315	0.369

within 2VPF and 3VPF systems enables the filter to follow fast and random variations of T motion compared with classic PF and VPF which seems to generate monotonous motion. Figure 2 shows also that our 3VPF system has the ability to follow even acute motions of the hand. In fact, optical flow prohibits particles from moving on motionless zones and adjust in some way their orientation. Then particles are further concentrated on moving objects. Thus, with adequate observation model and reference model, particles track the right target.

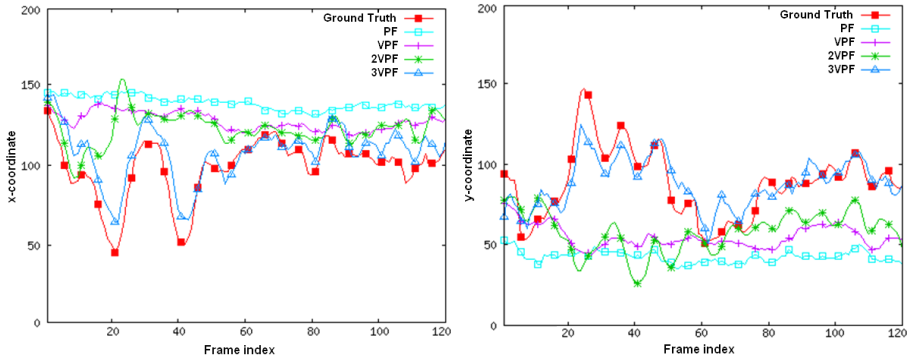

Fig. 2. x-coordinates and y-coordinates of the filter estimation along 120 frames of $\mathtt{S}_4$ (for the sake of clarity) for our four systems

5 Conclusion

We presented in this paper a hand tracking method based on a modified condensation algorithm and optical flow penalisation. The experiments done on an annotated database show the performance of our method compared to classic particle filter schemes. Nevertheless, we still have to improve our method so that it can handle multiple object tracking and occluded object.

References

1. Suchendra M. Bhandarkar and Xingzhi Luo. Integrated detection and tracking of multiple faces using particle filtering and optical flow-based elastic matching. *CVIU*, 113(6):708–725, June 2009.

2. Gary R. Bradski. Computer Vision Face Tracking For Use in a Perceptual User Interface. *Intel Technology Journal*, (Q2), 1998.
3. N. J. Gordon, D. J. Salmond, and A. F. M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, april 1993.
4. Michael Isard and Andrew Blake. Condensation - conditional density propagation for visual tracking. *Int J Comput Vison*, 29:5–28, 1998.
5. John KLEIN, Christele LECOMTE, and Pierre MICHE. Preceding car tracking using belief functions and a particle filter. In *IEEE ICPR - International Conference on Pattern Recognition*, pages 1–4, 2008.
6. Wei-Lwun Lu, Kenji Okuma, and James J. Little. Tracking and recognizing actions of multiple hockey players using the boosted particle filter. *Image Vision Comput.*, 27(1-2):189–205, jan 2009.
7. Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *Int Joint Conf Artif Intel*, volume 2 of *IJCAI'81*, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.
8. Carol Neidle, Stan Sclaroff, and Vassilis Athitsos. Signstream: A tool for linguistic and computer vision research on visual-gestural language data. *Behav Res Meth Ins C*, 33(3):311–320, 2001.
9. Patrick Perez, Jaco Vermaak, and Andrew Blake. Data fusion for visual tracking with particles. In *Proceedings of the IEEE*, pages 495–513, 2004.
10. Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, 2004.
11. Caifeng Shan, Tieniu Tan, and Yucheng Wei. Real-time hand tracking using a mean shift embedded particle filter. *PR*, 40(7):1958–1970, July 2007.
12. Leonid Sigal, Stan Sclaroff, and Vassilis Athitsos. Skin color-based video segmentation under time-varying illumination. *PAMI*, 26:862–877, 2003.
13. William C. Stokoe. Sign language structure: An outline of the visual communication systems of the american deaf. *Journal of Deaf Studies and Deaf Education*, 10(1), 2005.
14. Paul Viola and Michael J. Jones. Robust Real-Time face detection. *Int J Comput Vison*, 57(2):137–154, may 2004.
15. Shuying Yang, Weimin GE, and Zhang Cheng. Detecting and tracking moving targets on omnidirectional vision. *Transactions of Tianjin University*, 15(1):13–18, February 2009.
16. Angela Yao, Dominique Uebersax, Juergen Gall, and Luc Van Gool. Tracking people in broadcast sports. In *DAGM-PR*, pages 151–161, Berlin, Heidelberg, 2010. Springer-Verlag.